Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Feature aggregation and connectivity for object re-identification

Dongchen Han [a], Baodi Liu [b,*], Shuai Shao [b], Weifeng Liu [b], Yicong Zhou [c]

[a] College of Oceanography and Space Informatics, China University of Petroleum, Qingdao, 266580, China
[b] College of Control Science and Engineering, China University of Petroleum, Qingdao, 266580, China
[c] Department of Computer and Information Science, University of Macau, 999078, Macao Special Administrative Region of China

ARTICLE INFO

Keywords:
Object re-identification
Graph convolutional networks
Feature aggregation

ABSTRACT

In recent years, object re-identification (ReID) performance based on deep convolutional networks has reached a very high level and has seen outstanding progress. The existing methods merely focus on the robustness of features and classification accuracy but ignore the relationship among different features (i.e., the relationship between gallery–gallery pairs or probe–gallery pairs). In particular, a probe located at the decision boundary is the key to suppressing object ReID performance. We consider this probe as a hard sample. Recent studies have shown that Graph Convolutional Networks (GCN) significantly improve the relationship among features. However, applying the GCN to object ReID is still an open question. This paper proposes two learnable GCN modules: the Feature Aggregation Graph Convolutional Network (FA-GCN) and the Evaluation Connectivity Graph Convolutional Network (EC-GCN). Specifically, the pre-work selects an arbitrary feature extraction network to extract features in the object ReID dataset. Given a probe, FA-GCN aggregates neighboring nodes through the affinity graph of the gallery set. Afterward, EC-GCN uses a random probability gallery sampler to construct subgraphs for evaluating the connectivity of probe–gallery pairs. Finally, we jointly aggregate the node features and connectivity ratios as a new distance matrix. Experimental results on two person ReID datasets (Market-1501 and DukeMTMC-ReID) and one vehicle ReID dataset (VeRi-776) show that the proposed method achieves state-of-the-art performance.

1. Introduction

With the development of deep neural networks, the research interest in object re-identification (ReID) in the computer vision community, and the demand for intelligent video surveillance has increased significantly. Object ReID aims to capture objects with the same identity under different cameras. It usually searches for a specified object (i.e., a probe) in an image or video [1]. We need to find the images that match the object in the dataset (i.e., gallery set) and determine its identity. This object may appear at different times in the same or other locations. Object ReID is imperative and significant for intelligent systems, surveillance cameras, and public safety. For more details, please refer to Section 3.1.

It was recently, adopting deep neural networks as the backbone has become the optimal choice for ReID. Object ReID based on deep learning can be roughly divided into two research directions: (1) how to train a network to extract robust and linearly separable features [2]; (2) how to effectively retrieve identity through features given an object [3]. This paper is primarily concerned with the latter. The current general practice of ReID is to calculate the distance between a probe and the gallery set (i.e., cosine distance, Euclidean distance, and Mahalanobis distance), and use this distance as the basis for the ranking list. However, methods based on traditional metric learning cannot make full use of the training set when retrieving objects. In addition, object ReID will always have problems such as resolution, angle, and background under different cameras, resulting in too large differences in the same object. Even in a deep learning model, there will be a probe located on the decision boundary, which has a large impact on the k-nearest neighbors [4]. The task of retrieving objects is extremely destructive.

Currently, popular methods only focus on extracting robust features. When a feature point has many neighbors that do not match its class, we simply consider it a hard sample, as shown in Fig. 1(a). Hard samples are often difficult to train effectively by deep networks and loss functions and are unavoidable. Therefore, we introduce graph convolutional networks (GCN) to explore the connections between samples (nodes) and consider the relationship between the gallery–gallery pair or probe–gallery pair. Methods that existed before include: In [5], the

* Correspondence to: College of Control Science and Engineering, China University of Petroleum (East China), 66 Changjiang West Road, Huangdao District, Qingdao, Shandong, China.

E-mail addresses: hdc986600293@gmail.com (D. Han), thu.liubaodi@gmail.com (B. Liu), shuaishao@s.upc.edu.cn (S. Shao), liuwf@upc.edu.cn (W. Liu), yicongzhou@um.edu.mo (Y. Zhou).

https://doi.org/10.1016/j.patcog.2024.110869
Received 21 February 2023; Received in revised form 2 August 2024; Accepted 3 August 2024
Available online 8 August 2024

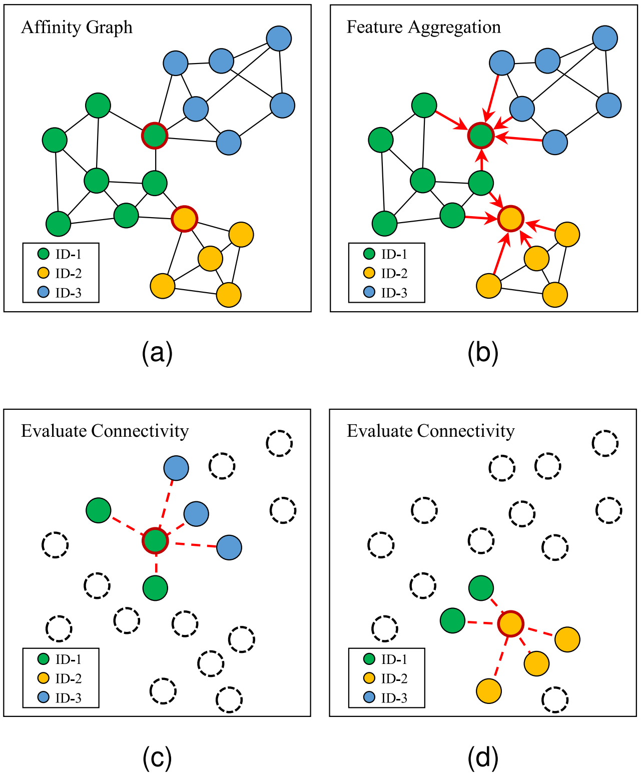


Fig. 1. Example of a method for handling a hard sample. The points enclosed by the red circle are hard samples. The same color represents the same identity. (a): construct the affinity graph using the features extracted by the backbone in the object ReID dataset. (b): take advantage of FA-GCN to aggregate features and information of neighboring nodes. (c-d): For each node, evaluate its connectivity with neighboring nodes by EC-GCN.

local features of the two images are regarded as a problem of image matching. Make full use of the alignment method to solve the occlusion problem. The Ref. [6] takes advantage of GCN to learn the topological structure of object features. Essentially, these two methods only learn the association between a single image or a pair of images and are not enough to distinguish representations. The Ref. [7] constructs a graph to represent the paired relationship between probe–gallery and uses this relationship to update the weights of GCN. The Ref. [8] uses the contextual information flow to predict the binary classification probability of each node to solve the hard sample problem. These two methods use probes to construct a subgraph and aggregate the context information of the nodes in the subgraph but only focus on local information and ignore global information.

Based on the above problems, we propose two joint modules, feature aggregation graph convolutional network (FA-GCN) and evaluate connectivity graph convolutional network (EC-GCN), for object re-identification. As illustrated in Fig. 1, for each sample (node) in the re-identification dataset, especially hard samples, our framework is divided into two angles to aggregate contextual information.

Specifically, FA-GCN first explores the structural relationship of the entire dataset and builds a graph with the number of nodes equal to the number of images in the dataset. Each node in the graph aggregates information about neighboring nodes. Through cross-entropy loss and contrastive loss, FA-GCN learns distinguishable node features and reclassifies each node.

Second, for hard samples at the edge of the class, the classification error is usually due to the small distance between the classes. To this end, EC-GCN constructs a subgraph for each probe. The size of the subgraph is the product of 1-hop and 2-hop. EC-GCN relies on the information provided by 2-hop nodes to 1-hop nodes to evaluate the probability of probe and 1-hop connection. While exploring the local subgraph information, since each node in FA-GCN aggregates neighbors' nodes, our subgraph actually contains a wealth of global information. Furthermore, we in EC-GCN introduce focal loss and Random Probability Gallery Sampler (RPGS) [9] to solve the problem of label imbalance in subgraphs and increase the diversity of graphs.

Third, We combine node features and connectivity probability to calculate the performance of object ReID.

To summarize, the contributions of the paper are as follows.

- We propose a novel framework containing two learnable GCN components, FA-GCN and EC-GCN. The framework can aggregate the information of the gallery set and solve the challenge of hard samples.
- Our framework is practical. Two components can be inserted into arbitrary deep neural networks and improve their performance.
- For the imbalance problem in the object ReID dataset, we formulate a new sampling strategy RPGS to increase the graph structure's diversity and combine focal loss to solve the issue of the ratio of positive and negative samples.
- We evaluate the proposed method on two person ReID benchmark datasets (Market-1501 [10] and DukeMTMC-reID [11,12]) and one vehicle ReID benchmark dataset (VeRi-776 [13,14]). The experimental results show that, especially in the vehicle ReID dataset, the proposed method achieves state-of-the-art performance. Based on the baseline, mAP and Rank-1 are significantly improved, indicating the great potential of FA-GCN and EC-GCN.

2. Related work

Our work is closely related to object ReID and Graph convolution networks. The following subsections summarize the literature on these two topics.

2.1. Object ReID

The research of object ReID is mainly attributed to the two fields of person ReID and vehicle ReID. Its challenges come from different low image resolutions, occlusions, lighting changes, complex camera environments, different viewpoints, etc. In addition, the demand for efficient retrieval speed and unforeseen scenarios has caused object ReID to be an unsolvable problem. A general ReID system requires four steps: data collection, data labeling, model training, and object retrieval. Researchers are mainly concerned with model training and object retrieval.

Model training aims to extract robust features. Early works mainly used hand-crafted features [15] to distinguish different objects by color, texture, shape, and spatiotemporal features. With the development of deep networks, ReID has been fully developed. Networks based on convolutional neural [16] and self-attention mechanisms [17] can extract powerful features from images or videos. He et al. [16] proposed foreground-aware pyramid reconstruction, which is used to calculate the score of occluded objects to solve the problem of object occlusion. Rao et al. [18] proposed to use counterfactual causality learning to interfere with the effect of visual attention on network prediction for adequate attention. He et al. [17] proposed a transformer-based framework that divides the image into a series of patches and shuffles the order to generate robust features, providing a solid baseline for ReID.

Object retrieval aims at generating accurate ranking lists. A common approach is to compute distances and ranks through metric learning. It relies on the feature space and retrieves by measuring the similarity between pictures. In the problem of object ReID, researchers need to measure the distance between a given object and a dataset. Usually, the closer the distance, the more similar it is, which is equivalent to the k-nearest neighbor algorithm. In general, the following measures

Table 1
Notations used in this paper.

Notations	Descriptions
$\mathbf{X}$	A training set.
$\mathbf{Q}$	A query set.
$\mathbf{G}$	A gallery set.
p	A node feature vector in the query set.
g	A node feature vector in the gallery set.
$\mathbf{F}$	The feature matrix
$\mathbf{F}^x$	The feature matrix of the train set.
$\mathbf{F}^p$	The feature matrix of the query set.
$\mathbf{F}^g$	The feature matrix of the gallery set.
$\mathcal{G}$	A graph.
$\mathbf{E}$	The set of nodes in a graph.
$\mathbf{V}$	The set of edges in a graph.
$\mathbf{A}$	The graph adjacency matrix.
$\mathbf{D}$	The graph degree matrix.
$\mathcal{G}(p)$	A subgraph sampled according to node p.
$\mathbf{V}(p)$	The set of nodes in $\mathcal{G}(p)$.
$\mathbf{E}(p)$	The set of edges in $\mathcal{G}(p)$.
$\mathbf{A}(p)$	A subgraph adjacency matrix sampled according to node p.
$\mathbf{Z}_l$	The node feature matrix of a graph in layer l.
$\mathbf{Z}(p)_l$	The node feature matrix of $\mathcal{G}(p)$ in layer l.
$\sigma(\cdot)$	The activation function.
$g(\cdot,\cdot)$	The mean aggregation operation.
$\mathbf{W}_l, \mathbf{B}$	Learnable model parameters.

are chosen to construct a ranking list. Euclidean distance measures the absolute distance of each point in space, which is directly related to the position coordinates of each point. Cosine distance measures the angle of space vectors, reflecting the difference in direction. Mahalanobis proposed the Mahalanobis distance, which can efficiently calculate the similarity between two groups of unknown samples. [19,20] use the method of multi-distance metric fusion to achieve great results. Recently, Zhong et al. [21] used Jaccard distance to measure the difference between two sets, and proposed a Re-Ranking ranking strategy. Zhang et al. [22] used the graph convolutional network GNN to reduce the excessive computational complexity of Re-Ranking to meet the needs of real-world applications.

2.2. Graph convolution network

Since traditional convolutional neural networks only process data in Euclidean space in the past few years, graph neural networks have developed rapidly.

Kipf et al. [23] proposed a method of graph convolutional neural network (GCN) applied to semi-supervised operations on graph structure data, which attracted strong attention. Wang et al. [24] developed a task of graph link prediction. The key idea is to infer the relationship between the input and its neighbors based on the context. Yang et al. [25] proposed a framework that combines graph convolutional networks with detection and segmentation, which significantly improves the performance of face clustering. Yang et al. [26] designed two graph convolutional networks, which are used to estimate the confidence of the node and the connected line of the edge, thereby improving the performance of the recognition model. Yang et al. [9] believe that the problem of imbalance has a huge impact on the GCN model based on link prediction, leading to a biased graph representation, so they proposed a reverse imbalance weighted sampling strategy.

Recently, the application of graph convolutional networks to ReID has also attracted much attention. Nguyen et al. [6] proposed to aggregate the attribute labels and visual features describing the person into a graph, take advantage of graph convolutional networks to learn the topological structure of the person, and integrate it into the ReID framework. To solve the problem of ReID of occluded people, Wang et al. [5] offered to learn topological information and high-order relationships to obtain features and robust alignment. Huang et al. [27] proposed a Reasoning and Tuning Graph Attention Network, which learns complete person representations of occluded images. To obtain robust visual similarity between images, Shen et al. [7] designed a Similarity-Guided graph neural network framework to calculate the relationship between probe and gallery. Ji et al. [8] proposed that the context-aware graph convolution network makes full use of the context information of ReID to solve the problem of hard samples and uses a hard gallery sampler to sample images, which can achieve higher performance.

3. Methodology

This section will introduce the five parts in detail: problem definition, framework overview, feature aggregation graph convolutional network, evaluation connectivity graph convolutional network, and final distance. The detailed descriptions of the notations could be found in Table 1.

3.1. Problem definition

Assume that we have a training set $\mathbf{X} = \{x_1, x_2, \ldots\}$, query set $\mathbf{Q} = \{p_1, p_2, \ldots\}$, and the gallery set $\mathbf{G} = \{g_1, g_2, \ldots\}$. We call each query set element a probe and define it as p, and the gallery set element as g. First, we must train the network $f(\cdot)$ on the training set. Then extract the features of datasets through this trained network. Usually, a feature can be expressed as $f^x = f(x)$. We define training set, query set, and gallery set feature matrices as $\mathbf{F}^x = \{f_1^x, f_2^x, \ldots\}$, $\mathbf{F}^p = \{f_1^p, f_2^p, \ldots\}$ and $\mathbf{F}^g = \{f_1^g, f_2^g, \ldots\}$, respectively. The goal of object ReID is to find images in the gallery set that match each probe in the probe set. It is a retrieval task. Our method is built upon extracted feature-based neural networks.

3.2. Framework overview

Our framework consists of four parts, namely feature extraction, FA-GCN, EC-GCN, and Rank, among which FA-GCN and EC-GCN are two learnable GCN modules for this task as shown in Fig. 2.

First, we extract features in the training set, query set, and gallery sets based on an arbitrary backbone. Specifically, we choose the backbone proposed in [28]. During training, we assume that each sample in the training set is a probe, and the rest represent the gallery set.

Second, the feature aggregation graph convolutional network (FA-GCN) processes all the extracted features. Features are constructed into affinity graphs using k-nearest neighbors (kNNs), and feature information of adjacent nodes is aggregated through L layer GCN and two loss functions (i.e., cross-entropy loss and contrastive loss).

Third, the ultimate goal of evaluating the connectivity graph convolutional network (EC-GCN) is to evaluate the probability of whether the probe is linked to the sampled gallery subset. We implement it through GCN and a binary classifier. We build a subgraph for each probe and input it into the GCN and binary classifier to obtain the connectivity. We need to be more careful about the local information in the subgraph. The hard gallery sampler proposed in [8] will cause the problem of imbalance between positive and negative samples. Therefore, we introduced focal loss [9,29] and a random probability gallery sampler (RPGS).

Finally, we obtain the FA-GCN ranking table according to the Mahalanobis distance through the node features output by FA-GCN. In EC-GCN, the connectivity of the probe–gallery pair is used as the distance to obtain the EC-GCN ranking list. Combine the distance matrix of these two modules to form the final ranking list. The final ranking list improves the effect of the backbone for object ReID. Moreover, the proposed framework can be applied to arbitrary neural networks.

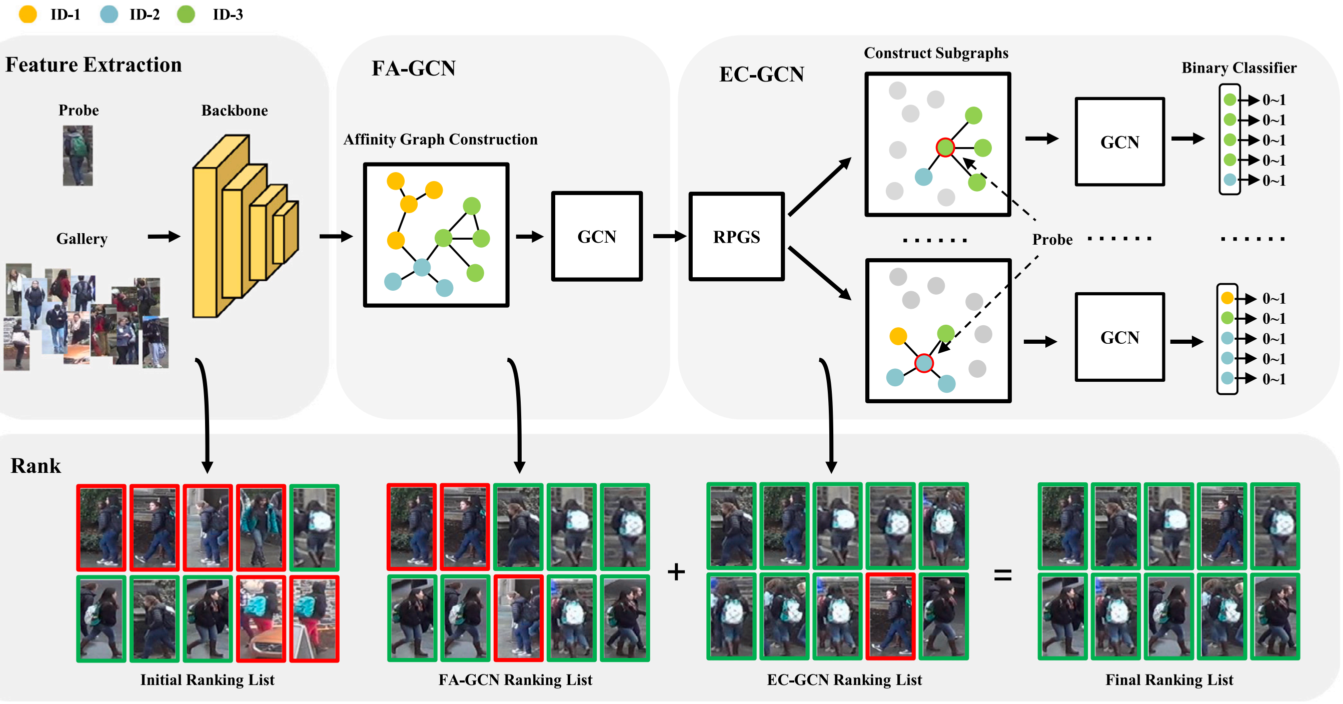


Fig. 2. Illustration of the proposed test framework for object ReID. It consists of four parts: Feature Extraction, FA-GCN, EC-GCN, and Rank. For specific procedures, please refer to Section 3.2.

3.3. Feature aggregation graph convolutional network

In this stage, we demonstrate how to implement FA-GCN. The main function of FA-GCN is to reclassify each node. A powerful graph convolutional network can aggregate features. The newly generated features aggregate information of neighboring nodes and can be better represented as context information. We adopt the features matrix $\mathbf{F} = \{f_1, f_2, \dots\} \in \mathbb{R}^{N\times d}$ of the dataset to construct a large affinity graph $\mathcal{G}(\mathbf{V}, \mathbf{E}, \mathbf{A})$, here $\mathbf{F} = \mathbf{F}^x$ in the training phase, $\mathbf{F} = \text{concat}[f^p, \mathbf{F}^g]$ in the testing phase, $\mathbf{V}$ is the node set of the graph, $\mathbf{E}$ is the edge set of the graph, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix of the graph, N is sum of the number of the features, and d is represents the dimension of a feature. Each element in $\mathbf{A}$ can be expressed as

$$\mathbf{A}_{i,j} = \begin{cases} \text{sim}(f_i, f_j), & if\, j \le k \\ 0, & if\, j > k, \end{cases} \tag{1}$$

where j is denoted as the index of the ith sample kNNs and $\text{sim}(\cdot,\cdot)$ is cosine similarity. For the similarity value of a row in $\mathbf{A}$, we keep the top kth. The FA-GCN consists of L layer GCN and linear layers, using $\mathbf{A}$ and feature matrix $\mathbf{F}$ as input. The calculation formula for each layer can be formulated as

$$\mathbf{Z}_{l+1} = \sigma(\text{concat}[\mathbf{Z}_l, \text{g}(\widetilde{\mathbf{A}}, \mathbf{Z}_l)]\mathbf{W}_l), \tag{2}$$

Where $\mathbf{Z}_l = \{z_1^l, z_2^l, \dots\} \in \mathbb{R}^{N\times d_{in}}$ is the node features in the l layer with $\mathbf{Z}_0$ is feature matrix $\mathbf{X}$ as the input of the first layer, d_{in} is the dimensions of the input features of each layer. $\widetilde{\mathbf{A}} = \mathbf{D}^{-1}(\mathbf{A}+\mathbf{I})$ is the normalized adjacency matrix. $\mathbf{D}^{-1} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{N'}(\mathbf{A}_i+\mathbf{I})_j$ and $\mathbf{D}_{ij} = 0$ if $i \neq j$. $\sigma(\cdot)$ is the ReLU activation function. $\text{g}(\cdot,\cdot)$ is the mean aggregation operation. $\mathbf{W}_l \in \mathbb{R}^{d'_{in}\times d_{out}}$ is a learnable weight matrix, where d_{out} is the dimension of the output features of each layer and d'_{in} is twice the dimension of the input feature. We employ a linear layer to predict the node classification for the last layer $\mathbf{Z}_L$. The linear layer can be formulated as

$$C_{cls} = \mathbf{Z}_L\mathbf{W}_{l+1} + \mathbf{B}, \tag{3}$$

where $C_{cls} \in \mathbb{R}^{N\times C}$ is the possibility of classification with C is the total number of classes, $\mathbf{W}_{l+1}$ is learnable parameters matrix, $\mathbf{B}$ is bias. The loss function is simply cross-entropy loss that can be formulated as

$$\mathcal{L}^{ce} = CrossEntropy(C_{cls}, y), \tag{4}$$

where y denotes the ground truth label. We apply supervised contrastive loss to obtain a more uniform high-dimensional feature space, which makes the sample features of the same label more compact to adapt to the re-identification task. The supervised contrastive loss can be written as

$$\mathcal{L}^{sup} = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i^{sup}, \tag{5}$$

$$\mathcal{L}_i^{sup} = \frac{1}{N_{y_i}}\sum_{j=1}^{N}1_{[y_i=y_j]}\log\frac{\exp\left(z_i\cdot z_j/\tau\right)}{\sum_{j=1}^{N}\exp\left(z_i\cdot z_j/\tau\right)}, \tag{6}$$

where N_{y_i} is the number of the same labels as the ith sample, and τ is a hyper-parameter. By weighting the sum of the two losses, we can get

$$\mathcal{L}^{fa} = \mathcal{L}^{ce} + \mathcal{L}^{sup}. \tag{7}$$

Usually, we need a weight to balance multiple losses. But in practice, the simple addition of the two losses has yielded effective performance.

3.4. Evaluation connectivity graph convolutional network

The general ReID uses Mahalanobis distance to obtain the distance matrix and then calculates the ranking list according to the distance matrix. This method can be summarized as the kNNs algorithm. As shown in Fig. 1(a), this distance is not necessarily an effective tool for measuring the matching degree of the probe–gallery pair. If the probe is a hard sample in a multi-class intersection area, the ranking list will have a lot of wrong-ordered images. Therefore, we use EC-GCN to predict whether there is a link between probe–gallery pairs. Allow the network to determine the connectivity between the probe–gallery pair by itself.

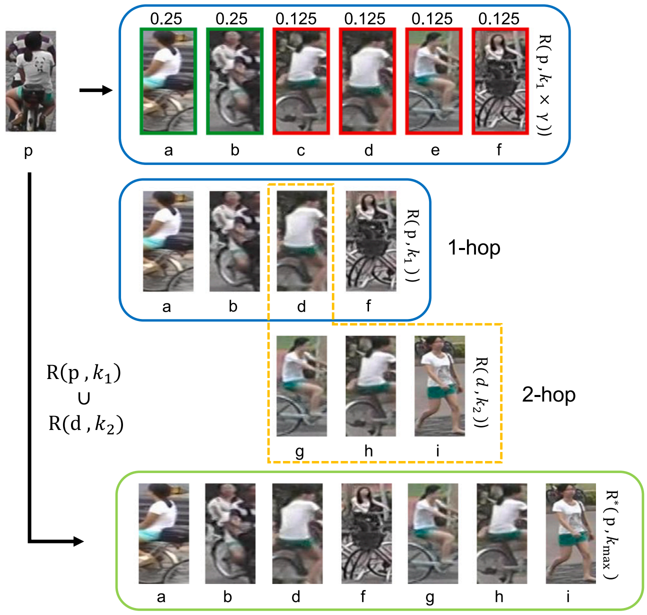


Fig. 3. Illustration of the Random Probability Gallery Sampler (RPGS). For a random probe P, we first sample $k_1 \times \gamma$ images from the gallery set. Then randomly select k images from these images as 1-hop according to the probability. Continue to sample 2-hop according to the image in 1-hop. Finally, we combine the images of 1-hop and 2-hop to form a sample set R^*. In this example, we set $k_1 = 4$, $k_2 = 3$, $\gamma = 1.5$.

3.4.1. *Hard gallery sampler*

In the EC-GCN stage, we use Section 3.3 method to re-extract features for the node. Use the Hard Gallery Sampler (HGS) proposed by [8] to sample the gallery set samples (i.e., n-hop neighbor). We only use 1-hop and 2-hop in this example, which are equivalent to probe–gallery search and gallery–gallery search. As shown in Fig. 3, for every element p from the probe set, we sample its k_1-nearest neighbors in the gallery set, that is, the k_1 images with the closest Euclidean distance as the 1-hop neighbors, and define it as $R(p,k_1) = \{g_1, g_2, \dots, g_{k_1}\}$. Then, For each g_i in $R(p,k_1)$ apply the same method to obtain $R(g_i,k_2)$ as the 2-hop neighbors. The final sample set can be calculated as

$$R^*(p,k_{max}) = R(p,k_1) \cup R(g_i,k_2), \tag{8}$$

where k_{max} is maximum number of sampled images.

3.4.2. *Random probability gallery sampler*

However, the n-hop sampling approach typically results in a significantly high proportion of positive samples within the graph. This imbalance between positive and negative instances can lead to a biased training process for Graph Convolutional Networks, where the model may tend to the identification of positive samples. This predisposition could undermine the network's ability to accuracy and generalization. Consequently, we propose RPGS, which sample $k_1 \times \gamma$ images before 1-hop, in Fig. 3, and define as $R(p, k_1 \times \gamma) = \{g_1, g_2, \dots, g_{k_1}, \dots, g_{k_1 \times \gamma}\}$. The green border represents the correct matching image, and the red represents the wrong. We add a sampling probability to each image in $R(p, k_1 \times \gamma)$, the calculation formula is

$$P_j = \begin{cases} \frac{1}{2} \times \frac{1}{N_p}, & g_j = p \\ \frac{1}{2} \times \frac{1}{N_n}, & g_j \neq p, \end{cases} \tag{9}$$

where j denote the index of every image in $R(p, k_1 \times \gamma)$, N_p is the number of positive samples, N_n is the number of negative samples. Randomly pick image from $R(p, k_1 \times \gamma)$ according to P_j until k_1 images are reached, forming our new $R(p,k_1)$ as 1-hop. Then follow the same method as before to sample 2-hop to form our final $R^*(p,k_{max})$.

3.4.3. *Construction subgraphs*

For each probe p, we take advantage of RPGS to sample the gallery image to obtain a new feature matrix $\mathbf{F}(p)$ and construct a subgraph $\mathcal{G}(p) = (\mathbf{V}(p), \mathbf{E}(p), \mathbf{A}(p))$. Use $\mathcal{G}(p)$ to explore the connectivity of the probe–gallery pair. Specifically, we adopt $R^*(p,k_{max})$ as node set $\mathbf{V}(p) = \{v_1, v_2, \dots, v_{k_{max+1}}\}$, and add edges among nodes to form an edge set $\mathbf{E}(p) \in \mathbb{R}^{|v| \times |v|}$, which $|v|$ is the total of the node. For a node $v_i \in \mathbf{V}(p)$, we look for its e-nearest neighbors (eNNs). If eNNs of v_i exist the same node v_j in $\mathbf{V}(p)$, connect them to form an edge $\mathbf{E}(p)(v_i, v_j) = 1$. Each value in the adjacency matrix $\mathbf{A}(p) \in \mathbb{R}^{k_{max} \times k_{max}}$ can be defined as

$$\mathbf{A}(p)_{i,j} = \begin{cases} 1, & if(\mathbf{E}(p)(v_i, v_j) = 1) \\ 0, & if(\mathbf{E}(p)(v_i, v_j) \neq 1), \end{cases} \tag{10}$$

Then we normalize all node features by subtracting the respective probe features in every $\mathcal{G}(p)$. Finally, EC-GCN consists of L' layer GCN and binary classifier, can be represented by

$$\mathbf{Z}(p)_{l'+1} = \sigma\left(BN\left(g\left(\tilde{\mathbf{A}}(p), \mathbf{Z}(p)_{l'}\right)\mathbf{W}_{l'}\right)\right), \tag{11}$$

where $\sigma(\cdot)$, $g(\cdot,\cdot)$, and $\mathbf{W}_{l'}$ are similar to Eq. (2), BN is the batchnorm layer, $\mathbf{Z}(p)_{l'}$ is the node feature matrix in the l' layer with the feature of $\mathbf{Z}(p)_0$ comes from FA-CCN, $\tilde{\mathbf{A}}(p) = \mathbf{D}^{-1}\mathbf{A}(p)$ with $\mathbf{D}_{ii} = \sum_{j=1} \mathbf{A}(p)_{ij}$ and $\mathbf{D}_{ij} = 0$ if $i \neq j$. For the binary classifier, we consist of three linear layers similar to Eq. (3) and two PReLU layers, of which the last linear layer is the binary classifier. Obtain the output of two classification values and normalize them to determine whether they are linked. Connectivity reflects the matching degree of each probe–gallery pair, which can be used for the distance matrix.

3.4.4. *Focal loss*

Inspired by [9,29], to solve the problem of serious imbalance between positive and negative samples in the sampling process. Too many negative samples cause positive samples to account for a small percentage of the loss. We define the last layer two classification output in EC-GCN as $\mathbf{P} = [p_p, p_n]$, where p_p is the probability of link, p_n is the probability of no link, $p_p + p_n = 1$. we use focal loss instead of cross-entropy loss, which can be formulated as

$$\mathcal{L}_{fc} = \begin{cases} -\alpha(1-p_p)^{\upsilon} \log p_p, & y_n = 1 \\ -(1-\alpha)(1-p_n)^{\upsilon} \log p_n, & y_n = 0, \end{cases} \tag{12}$$

where α is a hyper-parameter that balances positive and negative samples, usually 0.25, υ is the hard sample mining hyper-parameter. y is the true label of the two linked nodes.

3.5. *Final distance*

In the inference stage, we apply FA-GCN to obtain the aggregated features of each node as a new feature. Evaluate connectivity by EC-GCN, and use this probability as the feature distance. Therefore, our distance matrix can be defined as

$$d^*(p, g_i) = d(p, g_i) + \lambda d_g(p, g_i), \tag{13}$$

where p is a probe, $d(\cdot,\cdot)$ is the Mahalanobis distance calculation function, $d_g(\cdot,\cdot)$ is the distance matrix predicted by EC-GCN, λ is a hyper-parameter.

4. Experiments

In this section, we introduce three large-scale ReID benchmark datasets, including Market-1501 [10], DukeMTMC-reID [11,12], and VeRi-776 [13,14]. Then, a comparison with the state-of-the-art methods verifies the effectiveness of the two proposed modules. Finally, ablation experiments and visualization are performed to highlight some of the characteristics and properties of the proposed method.

Table 2
Comparisons with the state-of-the-art and baseline ReID methods on the Market-1501 and DukeMTMC-reID. The baseline and baseline* are from Ref. [28]. They use ResNet-50 [30] and ResNet50-ibn [31], respectively. RR denotes Re-Ranking [21].

Method	Reference	Market-1501			DukeMTMC-reID		
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
SGGNN [7]	ECCV2018	82.8	92.3	96.1	68.2	81.1	88.4
IDE+CamStyle [32]	TIP2019	68.7	88.1	-	53.4	75.2	–
FPR [16]	ICCV2019	86.5	95.4	-	72.3	76.0	–
ICT+CE [33]	TIP2020	83.7	94.2	-	73.5	86.6	–
HOReID [5]	CVPR2020	84.9	94.2	-	75.6	86.9	–
CAL [18]	ICCV2021	87.0	94.5	97.9	76.4	87.2	94.1
ADC+2O-IB [34]	CVPR2021	87.7	94.8	97.2	74.9	87.4	92.1
GPS [6]	CVPRW2021	87.8	95.2	98.4	78.7	88.2	95.2
TransReID [17]	ICCV2021	89.5	95.2	-	82.6	90.7	–
CAGCN [8]	AAAI2021	91.7	95.9	98.2	85.9	91.3	94.3
ALDER [35]	TIP2022	88.9	95.6	-	78.9	89.9	–
DFLN [36]	TCSVT2022	89.8	95.9	98.4	81.8	91.3	95.4
AGCL [37]	TPAMI2023	88.1	95.8	-	–	–	–
AdaSP [38]	CVPR2023	89.0	95.1	-	81.5	90.6	–
GlobalAP [39]	PR2023	90.1	96.0	-	82.0	91.2	–
baseline		85.9	94.5	98.2	76.4	86.5	93.9
baseline + FA-GCN		92.4	96.6	98.8	87.0	90.9	95.2
baseline + EC-GCN		93.7	96.0	98.3	88.2	91.1	95.2
baseline + FA-GCN + EC-GCN		94.1	96.8	98.6	88.6	91.5	95.3
baseline + RR		94.2	95.3	97.9	89.1	90.2	94.7
baseline + FA-GCN + EC-GCN + RR		95.3	96.7	98.5	89.9	91.1	95.1
baseline*		88.1	95.1	97.7	79.2	89.0	94.6
baseline*+ FA-GCN		93.2	96.2	98.6	88.1	92.2	95.7
baseline*+ EC-GCN		93.9	95.5	98.3	88.9	91.3	95.3
baseline*+ FA-GCN + EC-GCN		94.5	96.5	98.5	89.4	92.5	95.7

Table 3
Comparison with the state-of-the-art and baseline methods on the VeRi-776 dataset. The baseline is from Ref. [28]. The backbone is ResNet-50 [30]. RR denotes Re-Ranking [21].

Method	mAP	Rank-1	Rank-5
GRF+GGL [40]	61.7	89.4	95.0
SAVER [41]	79.6	96.4	98.6
CAL [18]	74.3	95.4	97.9
DSN [42]	76.3	94.8	97.5
CAGCN [8]	79.6	95.8	98.7
HRCN [43]	83.1	97.3	98.9
TransReID [17]	82.3	97.1	–
baseline	81.8	97.2	98.6
baseline + FA-GCN	84.1	97.5	98.8
baseline + EC-GCN	83.5	97.6	98.4
baseline + FA-GCN + EC-GCN	85.3	97.7	98.4
baseline + RR	83.9	97.0	98.7
baseline + FA-GCN + EC-GCN + RR	85.9	97.4	98.4

4.1. Datasets settings

We conduct experiments on two-person ReID datasets and one vehicle dataset.

Market-1501 [10] is a large-scale person ReID benchmark dataset consisting of 32,668 person rectangles taken under 6 camera lenses containing 1501 identities. In the training set, 12,936 person rectangles contain 751 identities, and each identity has about 17.2 images. A total of 19,732 person rectangles included 750 identities in the test set, and each identity has about 26.3 images. The test set is divided into two parts: query set and gallery set. The query set contains 3368 person rectangles drawn manually; the rest are gallery sets.

DukeMTMC-reID [11,12] is also a large-scale person ReID benchmark dataset, which comes from an 85-min high-resolution video taken under 8 camera lenses, with an image sampled every 120 frames, for a total of 36,411 images containing 1404 identities. All identities are equally divided into two parts; the first is the training set consisting of 16,522 images, and the other is the test set consisting of 19,889 images. The test set is subdivided into query and gallery sets, which have 2228 and 17,661 images, respectively.

VeRi-776 [13,14] is a large-scale vehicle ReID benchmark dataset, which comes from 20 cameras in a city area to shoot 24 h a day, a total of 45,622 images, containing 776 vehicles. The training set contains 37,778 images, the gallery set contains 11,579 images, and the query set includes 1678 images.

4.2. Evaluation metrics

Object ReID tasks usually adopt two evaluation metrics to evaluate the performance. The first one is Cumulative Matching Characteristics (CMC). Object ReID can be regarded as a sorting problem, and the hit probability of top-k calculated by the CMC curve can be an excellent measure of this problem. The other one is the mean average precision (mAP). This evaluation metric is often used for multi-label image classification and multi-target detection to measure the reliability of the model.

4.3. Experimental details

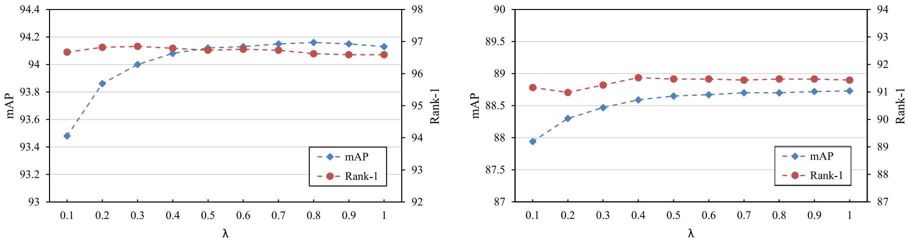
This section introduces the baseline model, parameter selection, and runtime experiment hardware and software.

4.3.1. Baseline model

For Makert-1501, DukeMTMC-reID, and VeRi-776 large benchmark dataset, we choose BoT-BS [28] as the baseline and extract all the features of the train set and test set. It utilizes ResNet-50 as the backbone, adding warm-up learning rate, label smoothing, BNNeck, last stride, center loss, and random erasing augmentation as a Bag of tricks, which significantly improves the accuracy of ReID. The baseline uses the official default configuration.

4.3.2. Parameter selection

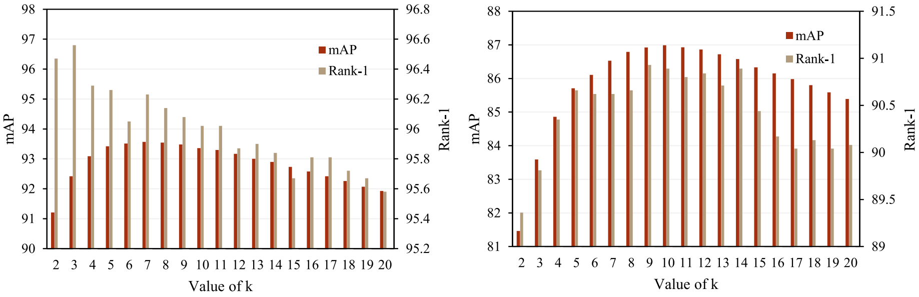
In the FA-GCN module, due to the difference in the number of images in different data sets, we choose different k values. In this experiment, the k of Makert-1501, DukeMTMC-reID, and VeRi-776 are 3, 10, and 20, respectively. We use one layer of GCN aggregation features and one layer of full connection layers as classifiers during the training phase. We train in 2000 epochs with the initialized learning rate $1e-4$ and attenuate 0.1 at 1000, 1600, and 1800 epochs.



(a) Market-1501

(b) DukeMTMC-reID

Fig. 4. The evaluation effect of different λ values in Eq. (13). (a) is Market-1501 dataset, and (b) is the dukeMTMC-reID dataset.



(a) Market-1501

(b) DukeMTMC-reID

Fig. 5. Evaluate the impact of mAP and rank-1 obtained by different k on Market-1501 and DukeMTMC-reID.

In the EC-GCN module, the RPGS method exists four hyper-parameters k_1, k_2, k_{max}, and γ. k_1 is the number of 1-hop. Since EC-GCN needs to determine the connectivity between probes and galleries in pairs, a larger value of k_1 will usually make the network a more powerful discriminative. As shown in Table 4, mAP and Rank-1 achieve their optimal values when $k_1 = 200$. Therefore, we use 200 as the default value in all experiments. k_2 is the number of 2-hop. Too large k_2 will cause the network to converge to unfavorable factors, so we set it to 10. k_{max} is the maximum number of nodes. We obtain by the formula $k_1 \times k_2$. γ is an important parameter to increase the number of graphs with different positive and negative sample ratios. As shown in Table 5, after adjusting the value of γ, the changes in mAP and Rank-1 achieve the overall optimal results around $\gamma = 1.5$. Therefore, we set it to 1.5. For the two hyper-parameters in focal loss, we set α to 0.25 and υ to 2. During the training phase, our framework consists of five layers of GCN and three layers of full connection, in which the dimension of the hidden layer is 4096. Our model is trained for 4 epochs, the initial learning rate is 0.01, and each epoch is reduced to 0.1% of the original. In Eq. (13), we set the value of λ to 0.4.

No special experiments require that our networks are all carried out under the same parameters.

4.3.3. Runtime experiment hardware and software

All our experiments are implemented on Linux servers. GPU is Tesla V100. We use python to train our model and Faiss to calculate k-nearest neighbors.

4.4. Comparison with state-of-the-arts

In this section, we divide object re-identification into two categories: person re-identification and vehicle re-identification. Compare them with the State-of-the-Arts methods.

4.4.1. Person re-identification

Compare our approach with baseline and other state-of-the-art person re-identification methods. The details of the comparison are shown in Table 2. Our method improves the performance with the addition of different modules.

On Market-1501, compared with the baseline without adding re-ranking, mAP, and Rank-1 of FA-GCN increased by 6.5% and 2.1%, respectively, EC-GCN increased by 7.8% and 1.5%, respectively, and FA-GCN plus EC-GCN increased by 8.2% and 2.3%, respectively. Compared with the state-of-the-art method, our method improves the mAP by 2.4%. When using Re-Ranking [21], the mAP increases to 94.2%.

On DukeMTMC-reID, compared with the baseline without re-ranking, FA-GCN increased by 10.6% mAP and 4.4% Rank-1, respectively, EC-GCN increased by 11.8% mAP and 4.6% Rank-1, respectively, and FA-GCN plus EC-GCN increased by 12.2% mAP and 5% Rank-1, respectively. Compared with the state-of-the-art method, our method improves the mAP by 2.7% mAP, reaching 88.6% mAP. Our method uses Re-Ranking to achieve 89.9% mAP.

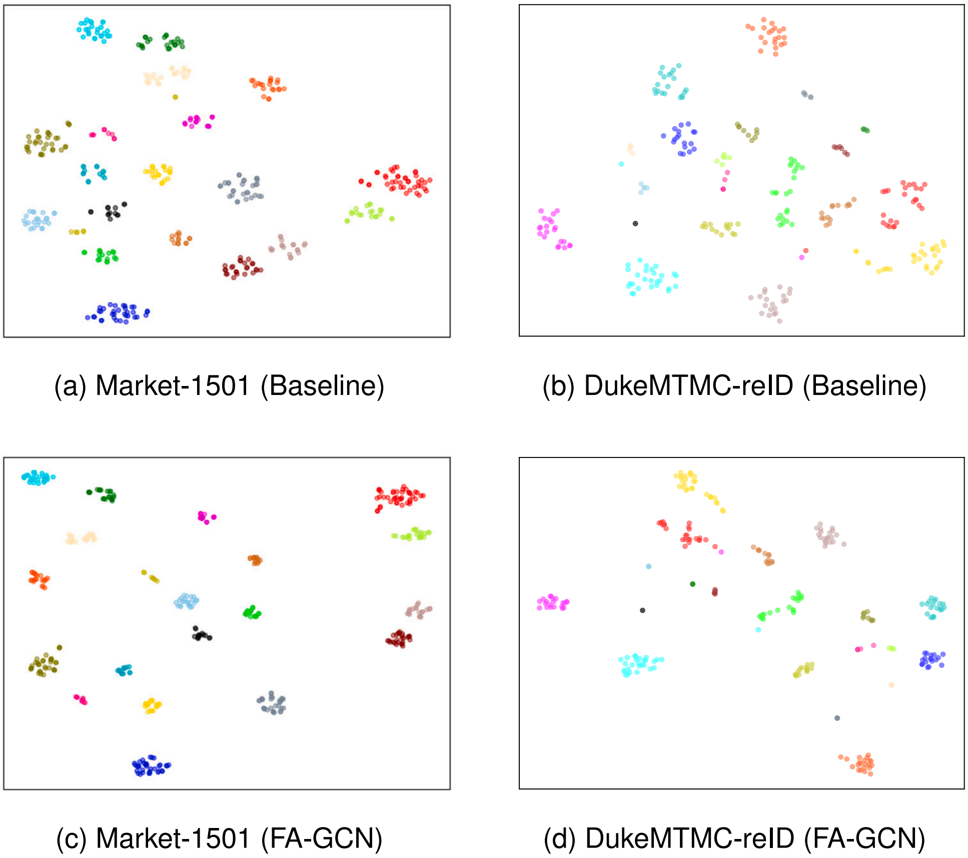
To verify the generality of the method, we conduct experiments on the network with the backbone of ResNet50-ibn [31]. FA-GCN plus EC-GCN achieves 94.5% mAP and 96.5% Rank-1 on Market-1501, and 89.4% mAP and 92.5% Rank-1 on DukeMTMC-reID. Compared with the baseline*, FA-GCN plus EC-GCN improve 6.4% mAP and 1.4% Rank-1 on Market-1501, 10.2% mAP, and 3.5% Rank-1 on DukeMTMC-reID.

4.4.2. Vehicle re-identification

As illustrated in Table 3, compared with the baseline on VeRi-776, FA-GCN reached 84.1% mAP and 97.5% Rank-1 and increased by 2.3% mAP and 0.3% Rank-1. EC-GCN reached 83.5% mAP and 97.6% Rank-1. FA-GCN plus EC-GCN achieved 85.3% mAP and 97.7% Rank-1 and increased by 3.5% mAP and 0.5% Rank-1. Compared with the latest network with the transformer added, our method improves mAP by 3%. Our approach is also applicable to vehicle ReID with Re-Ranking.

4.5. Ablation study

This section conducts four ablation experiments to describe our proposed framework. We specifically studied the influence of λ, k_1, and k on the different modules we suggested. In addition, the various performances produced by the paired combination of HGS, RPGS, CE loss, and focal loss are also shown.


(a) Market-1501 (Baseline)

(b) DukeMTMC-reID (Baseline)

(c) Market-1501 (FA-GCN)

(d) DukeMTMC-reID (FA-GCN)

Fig. 6. t-SNE visualization of 20 randomly sampled feature distributions on Market-1501 and DukeMTMC-reID. (a-b) comes from baseline [28]. (c-d) comes from FA-GCN. The same color indicates the same class.

Table 4
Evaluate the impact of different k_1 of RPGS on market-1501 and DukeMTMC-reID.

k_1	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
30	93.0	96.6	87.9	91.3
50	93.5	96.7	88.2	91.3
100	93.9	96.8	88.5	91.4
200	94.1	96.8	88.6	91.5

4.5.1. *Effects of value of λ*

In Fig. 4, we evaluated the effectiveness of different λ values on Market-1501 and DukeMTMC-reID. In (a) and (b), as the λ increases, the mAP of our FA-GCN plus EC-GCN also gradually increases. When $\lambda = 0.4$, it gradually stabilizes. Rank-1 is basically in a stable state. Therefore, we fixed λ to 0.4.

4.5.2. *Effects of value of k*

In the FA-GCN module, k determines the quantity and quality of aggregation for each node. As shown in Fig. 5, we report the mAP and the rank-1 for the value of k in the range of 2 to 20 on Makert-1501 and DukeMTMC-reID. We found that the value of k will show different performances with different datasets, and the performance is higher than the baseline in the tested range.

4.5.3. *Effects of value of k_1*

In Table 4, we describe in detail the effects of k_1 values of 30, 50, 100, and 200. The performance of FA-GCN plus EC-GCN gradually increases with the value of k_1. We speculate that it may be because a large subgraph is generated with the increase of k_1. The network needs to judge the connectivity of the nodes in each subgraph, which indirectly improves the robustness of GCN.

Table 5
Evaluate the impact of different γ of RPGS on market-1501 and DukeMTMC-reID.

γ	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
1	93.9	96.6	88.3	91.1
1.5	94.1	96.8	88.6	91.5
2	94	96.8	88.5	91.3
3	93.9	96.7	88.5	91.2

Table 6
The ablation study of different components on the Market-1501 dataset.

case	HGS	RPGS	CE loss	focal loss	mAP	Rank-1
(a)	✓		✓		93.3	96.5
(b)		✓	✓		93.5	96.7
(c)	✓			✓	93.9	96.6
(d)		✓		✓	94.1	96.8

4.5.4. *Effects of value of γ*

In Table 5, We conducted experiments on the Market-1501 and DukeMTMC-reID datasets to evaluate the impact of different values of γ, within the range of 1 to 3. As shown in the table, The results indicate that both mAP and rank-1 initially increase and then decrease as γ increases. Overall, the optimal value is achieved at $\gamma = 1.5$.

4.5.5. *Comparison of the different methods*

To verify the effectiveness of the methods, in Table 6, we combined each component in pairs and observed its performance. Case (a) represents a base performance. Case (b) tests the performance of RPGS. Case (c) tests the performance of focal loss. Case (d) is a combination of RPGS and focal loss. Compared with case (a), case (d) has increased mAP by 0.8% and rank-1 by 0.3%, validating the necessity of the two components.

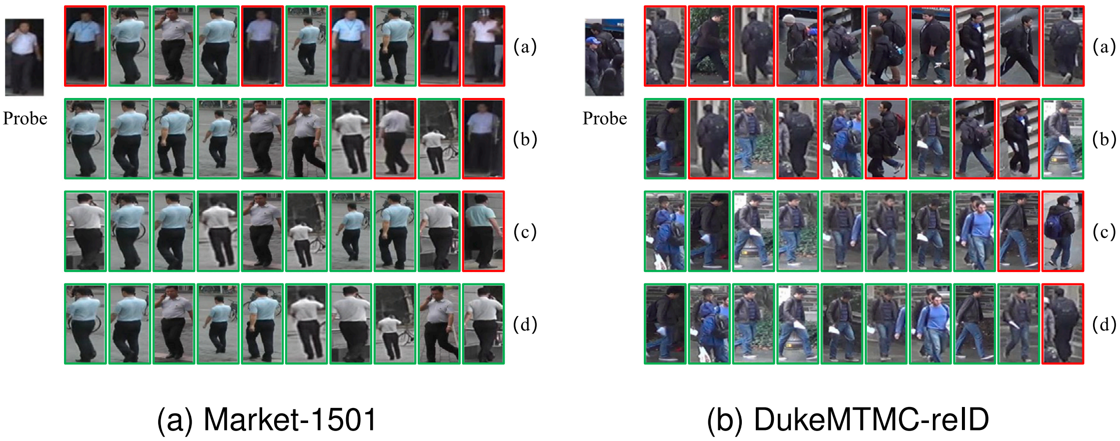


Fig. 7. Compare the visual results of the identity images retrieved by the baseline and our method on Market-1501 and DukeMTMC-reID, respectively. (a) represents the model proposed in [28] as our baseline. (b) is the result of FA-GCN. (c) is the result of EC-GCN. (d) is the result of FA-GCN plus EC-GCN. The four lines corresponding to a probe are the 10 results searched in the order of distance in four different methods. The green box matches a probe, but the red box does not.

4.6. *Visualization*

4.6.1. *t-SNE*

In this work, we visualize the t-SNE graphs of the two datasets (i.e., Market-1501 and DukeMTMC-reID), as shown in Fig. 6. Compared with the baseline, our method can significantly reduce the intra-class distance and increase the inter-class distance. It has a corrective effect on the difficult samples of the class edge. After that, we use EC-GCN to optimize further hard samples that are still indistinguishable and eliminate the challenge of hard samples while obtaining global information.

4.6.2. *Ranking list*

As shown in Fig. 7, we offer the visualization results of images retrieved by two probes in four different methods. (a) displays the images retrieved by the baseline. (b-d) displays the images retrieved by FA-GCN, EC-GCN, and a combination of the two methods. Our method retrieves more correct images than the baseline, can achieve a higher accuracy rate, and help the model improve performance.

5. Conclusion

This paper proposes two graph convolutional network modules, FA-GCN and EC-GCN, to be applied to object re-identification. FA-GCN can effectively aggregate the context information of neighbor nodes so that nodes with the same identity can reduce the distance and re-extract features. EC-GCN processes hard samples that are still inseparable after FA-GCN and evaluates the connectivity with neighbor nodes by a binary classifier. We replace distance with connectivity to reduce the time loss caused by feature retrieval. Experiments prove the performance of EC-GCN alone.

The advantages of this work are: By introducing the FA-GCN and EC-GCN modules, we effectively address the hard sample problem in object re-identification. This method not only improves feature discrimination but also enhances the connectivity between the probe and gallery. It can be integrated into any deep neural network, boosting performance with great practicality and flexibility. Despite its advantages in ReID, this work still has some limitations: (1) The introduction of two GCN modules increases the model's complexity and computational cost, which might require more computing resources in practical applications. (2) The performance of the method relies on the quality and diversity of the training data.

This research provides two directions for future work: (1) Further optimize the structures of FA-GCN and EC-GCN to reduce computational complexity while enhancing the model's robustness and interpretability. (2) Explore domain adaptation across different datasets to improve the generalizability of the method in various environments and scenarios.

CRediT authorship contribution statement

Dongchen Han: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Baodi Liu:** Supervision, Resources, Funding acquisition, Conceptualization. **Shuai Shao:** Visualization, Validation, Software, Project administration, Methodology. **Weifeng Liu:** Validation, Supervision, Investigation, Funding acquisition, Conceptualization. **Yicong Zhou:** Validation, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 61671480, the National Natural Science Foundation of China under Grant 62372468, the Natural Science Foundation of Shandong Province, China, under Grant ZR2019MF073, the Natural Science Foundation of Shandong Province, China, under Grant ZR2023MF008, the Qingdao Natural Science Foundation under Grant 23-2-1-161-zyyd-jch, the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008, the Major Basic Research Projects in Shandong Province under Grant ZR2023ZD32.

References

[1] B. Jiao, L. Yang, L. Gao, P. Wang, S. Zhang, Y. Zhang, Vehicle re-identification in aerial images and videos: Dataset and approach, IEEE Trans. Circuits Syst. Video Technol. (2023) http://dx.doi.org/10.1109/tcsvt.2023.3298788.

[2] Y. Xie, H. Wu, Y. Lin, J. Zhu, H. Zeng, Pairwise difference relational distillation for object re-identification, Pattern Recognit. 152 (2024) 110455, http://dx.doi.org/10.1016/j.patcog.2024.110455.

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: Automatic query expansion with a generative feature model for object retrieval, in: ICCV, IEEE, 2007, pp. 1–8, http://dx.doi.org/10.1109/ICCV.2007.4408891.

[4] D. Han, S. Shao, W. Liu, B.-D. Liu, Object re-identification with distribution corrected ranking list, Neurocomputing 506 (2022) 117–127, http://dx.doi.org/10.1016/j.neucom.2022.07.062.

[5] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: Learning relation and topology for occluded person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6449–6458, http://dx.doi.org/10.1109/CVPR42600.2020.00648.

[6] B.X. Nguyen, B.D. Nguyen, T. Do, E. Tjiputra, Q.D. Tran, A. Nguyen, Graph-based person signature for person re-identifications, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3492–3501, http://dx.doi.org/10.1109/cvprw53098.2021.00388.

[7] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 486–504, http://dx.doi.org/10.1007/978-3-030-01267-0_30.

[8] D. Ji, H. Wang, H. Hu, W. Gan, W. Wu, J. Yan, Context-aware graph convolution network for target re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 1646–1654, http://dx.doi.org/10.1609/aaai.v35i2.16257.

[9] H. Yang, X. Chen, F. Zhang, G. Hei, Y. Wang, R. Du, GCN-based linkage prediction for face clusteringon imbalanced datasets: An empirical study, 2021, arXiv preprint arXiv:2107.02477.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124, http://dx.doi.org/10.1109/ICCV.2015.133.

[11] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, http://dx.doi.org/10.1109/ICCV.2017.405.

[12] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision Workshop on Benchmarking Multi-Target Tracking, 2016, http://dx.doi.org/10.1007/978-3-319-48881-3_2.

[13] X. Liu, W. Liu, T. Mei, H. Ma, A deep learning-based approach to progressive vehicle re-identification for urban surveillance, in: European Conference on Computer Vision, Springer, 2016, pp. 869–884, http://dx.doi.org/10.1007/978-3-319-46475-6_53.

[14] X. Liu, W. Liu, T. Mei, H. Ma, Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance, IEEE Trans. Multimed. 20 (3) (2017) 645–658, http://dx.doi.org/10.1109/TMM.2017.2751966.

[15] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, Springer, 2008, pp. 262–275, http://dx.doi.org/10.1007/978-3-540-88682-2_21.

[16] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, J. Feng, Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8450–8459, http://dx.doi.org/10.1109/ICCV.2019.00854.

[17] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15013–15022, http://dx.doi.org/10.1109/iccv48922.2021.01474.

[18] Y. Rao, G. Chen, J. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1025–1034, http://dx.doi.org/10.1109/ICCV48922.2021.00106.

[19] S. Zhang, J. Huang, H. Li, D.N. Metaxas, Automatic image annotation and retrieval using group sparsity, IEEE Trans. Syst. Man Cybern. 42 (3) (2012) 838–849, http://dx.doi.org/10.1109/TSMCB.2011.2179533.

[20] S. Zhang, M. Yang, T. Cour, K. Yu, D.N. Metaxas, Query specific rank fusion for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2014) 803–815, http://dx.doi.org/10.1109/TPAMI.2014.2346201.

[21] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327, http://dx.doi.org/10.1109/cvpr.2017.389.

[22] X. Zhang, M. Jiang, Z. Zheng, X. Tan, E. Ding, Y. Yang, Understanding image retrieval re-ranking: a graph neural network perspective, 2020, arXiv preprint arXiv:2012.07620.

[23] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[24] Z. Wang, L. Zheng, Y. Li, S. Wang, Linkage based face clustering via graph convolution network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1117–1125, http://dx.doi.org/10.1109/CVPR.2019.00121.

[25] L. Yang, X. Zhan, D. Chen, J. Yan, C.C. Loy, D. Lin, Learning to cluster faces on an affinity graph, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2298–2306, http://dx.doi.org/10.1109/cvpr.2019.00240.

[26] L. Yang, D. Chen, X. Zhan, R. Zhao, C.C. Loy, D. Lin, Learning to cluster faces via confidence and connectivity estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13369–13378, http://dx.doi.org/10.1109/CVPR42600.2020.01338.

[27] M. Huang, C. Hou, Q. Yang, Z. Wang, Reasoning and tuning: Graph attention network for occluded person re-identification, IEEE Trans. Image Process. 32 (2023) 1568–1582, http://dx.doi.org/10.1109/tip.2023.3247159.

[28] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, http://dx.doi.org/10.1109/CVPRW.2019.00190.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988, http://dx.doi.org/10.1109/iccv.2017.324.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[31] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: Enhancing learning and generalization capacities via ibn-net, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 464–479, http://dx.doi.org/10.1007/978-3-030-01225-0_29.

[32] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019) 1176–1190, http://dx.doi.org/10.1109/TIP.2018.2874313.

[33] F. Xu, B. Ma, H. Chang, S. Shan, Isosceles constraints for person re-identification, IEEE Trans. Image Process. 29 (2020) 8930–8943, http://dx.doi.org/10.1109/TIP.2020.3020648.

[34] A. Zhang, Y. Gao, Y. Niu, W. Liu, Y. Zhou, Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 598–607, http://dx.doi.org/10.1109/CVPR46437.2021.00066.

[35] Q. Zhang, J. Lai, Z. Feng, X. Xie, Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification, IEEE Trans. Image Process. 31 (2022) 352–365, http://dx.doi.org/10.1109/TIP.2021.3128330.

[36] S. Yang, W. Liu, Y. Yu, H. Hu, D. Chen, T. Su, Diverse feature learning network with attention suppression and part level background suppression for person re-identification, IEEE Trans. Circuits Syst. Video Technol. 33 (1) (2022) 283–297, http://dx.doi.org/10.1109/tcsvt.2022.3199394.

[37] H. Zhang, M. Liu, Y. Li, M. Yan, Z. Gao, X. Chang, L. Nie, Attribute-guided collaborative learning for partial person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 14144–14160, http://dx.doi.org/10.1109/tpami.2023.3312302.

[38] X. Zhou, Y. Zhong, Z. Cheng, F. Liang, L. Ma, Adaptive sparse pairwise loss for object re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19691–19701, http://dx.doi.org/10.1109/cvpr52729.2023.01886.

[39] Y. Liu, Y. Liang, P. Wang, Z. Chen, C. Ding, GlobalAP: Global average precision optimization for person re-identification, Pattern Recognit. 142 (2023) 109682, http://dx.doi.org/10.1016/j.patcog.2023.109682.

[40] X. Liu, S. Zhang, X. Wang, R. Hong, Q. Tian, Group-group loss-based global-regional feature learning for vehicle re-identification, IEEE Trans. Image Process. 29 (2020) 2638–2652, http://dx.doi.org/10.1109/TIP.2019.2950796.

[41] P. Khorramshahi, N. Peri, J.-c. Chen, R. Chellappa, The devil is in the details: Self-supervised attention for vehicle re-identification, in: European Conference on Computer Vision, Springer, 2020, pp. 369–386, http://dx.doi.org/10.1007/978-3-030-58568-6_22.

[42] W. Zhu, Z. Wang, X. Wang, R. Hu, H. Liu, C. Liu, C. Wang, D. Li, A dual self-attention mechanism for vehicle re-identification, Pattern Recognit. 137 (2023) 109258, http://dx.doi.org/10.1016/j.patcog.2022.109258.

[43] J. Zhao, Y. Zhao, J. Li, K. Yan, Y. Tian, Heterogeneous relational complement for vehicle re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 205–214, http://dx.doi.org/10.1109/iccv48922.2021.00027.

Dongchen Han received the B.S. degree from the Department of Information Engineering, Shandong Youth University for Political Sciences, Shandong, China, in 2020. He is currently pursuing the M.S. degree in college of oceanographer and space informatics with the China University of Petroleum (East China), Qingdao, China. His research interests include object re-identification and contrastive learning.

Baodi Liu (IEEE Member, ACM Member) is currently an Associate Professor with the College of Control Science and Engineering, China University of Petroleum (East China), China. He received the B.S. degree in Signal and Information Processing from China University of Petroleum (East China) in 2007. He got the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University in 2013. He was a Visiting Scholar at University of California, Merced, from 2019 to 2020. His research interest includes Image Processing, Computer Vision, and Machine Learning.

Shuai Shao received his M.S. degree in College of Control Science and Engineering, China University of Petroleum (East China). Currently, he is pursuing his Ph.D. degree in College of Control Science and Engineering, China University of Petroleum (East China). In Ph.D career, he served as a research assistant in Tsinghua University (2019–2020) and has published 9 papers, including ACMMMM, ICME (oral) et.al.

Weifeng Liu received the double B.S. degrees in automation and business administration and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2002 and 2007, respectively. He was a Visiting Scholar with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia, from 2011 to 2012. He is currently a Full Professor with the College of Information and Control Engineering, China University of Petroleum, Qingdao, China. He has authored or coauthored a dozen papers in top journals and prestigious conferences, including four Essential Science Indicators (ESI) highly cited papers and two ESI hot papers. His research interests include computer vision, pattern recognition, and machine learning. Prof. Liu serves as an Associate Editor for the Neural Processing Letters, the Co-Chair for the IEEE SMC Technical Committee on Cognitive Computing, and a Guest Editor for special issue of the Signal Processing, the IET Computer Vision, the Neurocomputing, and the Remote Sensing. He also serves over 20 journals and over 40 conferences.

Yicong Zhou received the B.S. degree from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees from Tufts University, Massachusetts, USA, all in electrical engineering. He is currently a Professor and Director of the Vision and Image Processing Laboratory in the Department of Computer and Information Science at University of Macau. His research interests include image processing, computer vision, machine learning, and multimedia security. Dr. Zhou is a Fellow of the International Society for Optical Engineering (SPIE), and a Senior Member of the IEEE and CCF (China Computer Federation). He was a recipient of the Third Price of Macao Natural Science Award in 2014 and 2020. He is a Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He serves as an Associate Editor for IEEE Transactions on Neutral Networks and Learning Systems (TNNLS), IEEE Transactions on Cybernetics (TCYB), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on Geoscience and Remote Sensing (TGRS), and four other journals. He was listed as “World’s 2% Scientists” on the Stanford University Releases List 2020, 2021 and the “Highly Cited Researcher” in the Web of Science 2020, 2021.